LANGUAGE TESTING
RESEARCH COLLOQUIUM

JOINTLY HOSTED BY THE
DEFENSE LANGUAGE INSTITUTE
AND THE
MONTEREY INSTITUTE OF INTERNATIONAL STUDIES

FEBRUARY 27, 28 & MARCH 1, 1986

ORGANIZING COMMITTEE:

KATHLEEN M. BAILEY, MIIS, CO-CHAIR
RAY CLIFFORD, DLI, CO-CHAIR
TED CALE, MIIS, ADMINISTRATIVE ASSISTANT
GARY DUDNEY, CTB McGRAW HILL
SUZANNE FOSTER, PROTOCOL OFFICE, DLI
PIERRETTE HARTER, PROTOCOL OFFICE, DLI
MARTHA HERZOG, TESTS AND STANDARDS DIVISION, DLI
Thursday, February 27

8:45  Opening Remarks
      Kathi Bailey and Ray Clifford

9:00  Testing the Receptive Skills: Listening Comprehension
      A Workshop on DLI Testing
      Herbert Davy, Martha Herzog, and Ellen Mitchell

11:30 No-host Lunch at the DLIFLC Officers’ Club

1:00  Testing the Receptive Skills: Reading Comprehension
      A Workshop on DLIFLC Testing
      Herbert Davy, Martha Herzog, and Ellen Mitchell

4:15  Concluding Remarks and Focus for the Colloquium
      Ray Clifford

Friday, February 28

8:45  Opening Remarks
      Kathi Bailey

9:00  Development of the Strategy Inventory for Language Learning
      Rebecca L. Oxford

9:30  Testing the Transfer of ACTFL/ETS/ILR Reading Proficiency Scales to Academia
      Dale Lange and Pardee Lowe, Jr.

10:15 Investigations Into the Appropriate Use of Various Item Types for Testing Specific Levels of Reading Proficiency
      Sandra S. McIntyre

10:45 An Evaluation Scale for Integrative Writing Tasks on a Performance Test of English for Academic Purposes
      Stan Jones, Ellen Cray, Andi Gray, and Linda Librande

11:15 Performance Profiles for Academic Writing
      Liz Hamp-Lyons and Peter Hargreaves

11:45 Scalar Analysis of Composition Ratings
      Grant Henning and Fred Davidson

12:15 Lunch at the Outrigger; afternoon free for sightseeing, visiting the Monterey Bay Aquarium, or listening to additional presentations.
Saturday, March 1, 1986

9:00 A Study of the Comparability of Speaking Proficiency Interview Ratings Across Three Government Agencies
   John L.D. Clark

9:30 Testing Oral Second Language Proficiency in the Canadian Government
   Judy A. Purdom

10:00 The Graphic Representation of Language Competence: Mapping EFL Proficiency Using Multidimensional Scaling Techniques
   Russanne Hozayin

10:45 A Comparison of Two Forms of Tutor Assessment
   Peter Hargreaves

11:15 An Experiment with an On-Line Narrative Discourse Test
   Steven Ross

11:45 Ten Parameters for Measuring Pronunciation
   Garry Molholt

12:15 Lunch in Munakata Hall

1:15 Innovation through Computer-Based Language Testing (CBELT)
   Charles Alderson

1:45 Some Problems Concerning the Interpretation of Passage Correction Tests
   Terence Odlin

2:15 Utilizing Rasch Analysis to Detect Cheating on Language Exams
   Harold Madsen

2:45 Coffee Break

3:00 Round Table Discussion Groups

4:00 Group Summaries

5:00 Concluding Remarks and Business Meeting
   Kathi Bailey and Ray Clifford

6:00 No-host cocktails at "Two Guys from Italy" (near the Fairgrounds Travelodge); meet for carpooling

7:30 Supper at Ray Clifford's home
Development of the Strategy Inventory for Language Learning

This paper describes the development of a new, comprehensive second-language (L2) learning strategy measurement scale known as the "Strategy Inventory for Language Learning (SILL)." Learning strategies are steps taken by the learner to facilitate the acquisition, storage, retrieval, or use of information (O'Malley, Russo, & Chamot, 1983). Only in the last decade has a sufficiently strong case been made for the importance of learning strategies and for the learner's responsibility for furthering his or her own progress. Recently, learning strategies and other attributes of the learner have begun to gain the measurement attention they deserve. The military services have been at the forefront of much of the recent measurement work concerning learning strategies, particularly in the L2 area. The measurement scale described in this paper was developed for the Army for later use in a major, multi-year study on language skill change (especially loss) after training is completed. The study will be conducted under the auspices of the Army Research Institute and the Defense Language Institute Foreign Language Center. Additionally, the SILL will be used in another study of language loss sponsored by the National Security Agency and conducted by the Center for Applied Linguistics and the University of Pennsylvania.

The SILL is a 135-item survey that assesses how individuals use learning strategies during and after language training. It takes 30-45 minutes to administer and is geared toward persons studying a second language. Previous learning strategy scales have evidenced a number of problems, such as inconsistent strategy definitions, low subscale reliability, response "fakability," lack of empirical validation of good strategies, fragmentary approaches, and limited diagnostic capability. The SILL is an attempt to correct as many of these problems as possible. This paper explains the methodology used to develop and revise the SILL. Steps involved in the development process included a thorough review of the research literature on learning strategies, particularly on L2 learning strategies; creation of a 220-item prototype scale; clinical trials of that version; revision based on results of the clinical trials and expert judgments; a 500-person pilot test involving all four military services; factor analysis of the pilot test data; correlations of the pilot test version of the SILL with biographical data and DLIFLC proficiency test data; final reliability and validity statistical analyses; and use of the instrument in the studies named above. The literature review included development of a new language learning strategy taxonomy, from which the SILL takes its theoretical orientation.
Testing the Transfer of ACTFL/ETS/ILR Reading Proficiency Scales to Academia

In its ACTFL/ETS version the ILR oral proficiency scale and accompanying levels have been successfully transferred to academia. On the other hand, the reading proficiency scale, even in its ACTFL/ETS form, has been less well received.

Academics consistently comment that the reading scale seems harder to grasp. Moreover, they understand the rationale for using it less fully, and question both whether reading proficiency test performances in academia could be rated accurately according to the scale and particularly whether passages, the comprehension of which forms the basis for ratings, could be properly graded for level.

This paper reports on a study comprising two tasks connected with the pivotal final task--assigning levels to passages. The tasks were specifically designed to suggest that the ACTFL/ETS and ILR scales provide a meaningful basis for rank-ordering and assigning levels. In the study, English passages carefully rank-ordered and graded for level by the ILR Testing Committee were given blind and in random order to 25 subjects. In the first task the subjects were asked to rank-order the passages, in the second task to assign to the same passages suitable ACTFL/ETS and ultimately ILR levels. Subjects were a group of test developers connected with the University of Minnesota’s project to write entrance and exit proficiency tests in French, German, and Spanish for the new foreign language proficiency requirement in the University of Minnesota’s College of Liberal Arts.

Already familiar with the ACTFL/ETS scales in all skills, the group was asked after a short review of both the ACTFL/ETS and the ILR versions to rank a series of passages in order of difficulty. Next the group was asked to assign ACTFL designations to the same passages. Both tasks were undertaken on the first day of a five day workshop during which items for testing all four skill modalities were to be written. Without further formal review of the scales the group proceeded to write suitable proficiency test items under the guidance of experienced test developers. On the last day the two tasks from the first day were repeated.

The study reports on the extent of agreement of these two tasks between the initial and final rank-ordering and assignment of levels suggested by the group and the ordering and levels assigned originally by the ILR testing committee. The success attained on the two tasks suggests strongly that the ILR reading proficiency scale can probably be transferred to academia as successfully as the oral scale has been.
Investigations Into the Appropriate Use of Various Item Types for Testing Specific Levels of Reading Proficiency

The purpose of this study was to investigate the appropriate use of various formats for testing specific levels of reading proficiency. Item response data concerning various item formats have been collected at the Defense Language Institute through the production and use of the Defense Language Proficiency Tests (DLPT IIIs). Item formats have included the identification of signs, identification of underlined information, multiple choice and True-False-Not Addressed items, multiple choice comprehension questions, and cloze completions. (Examples will be provided.) The DLPT IIIs, which have been developed for several languages, are written according to a design which specifies increasing difficulty of items ranging from 0+ to 3 on the Interagency Language Scale of Language Proficiency.

The study examined performance as measured by DLPT III tests in four languages: French, Spanish, Korean and Chinese. The subjects consisted of language students at DLI, students at universities, and military and diplomatic linguists in the field. The data from an average of 76 subjects were used as the basis for examining performance on each of the four languages. Subjects were tested in group sessions where they were given a test booklet and an answer-sheet. They were given two and one-half hours to complete the test.

Item response curve analyses of the DLPT III reading tests yielded information concerning the power of discrimination and relative difficulty level of the item formats. Proficiency level of the students (as measured by the reading section of the oral interview) was plotted on the Y axis. The resulting slopes indicated that identification of signs and information identification items discriminated most highly between levels 0+ and 1. Multiple choice comprehension questions discriminated between levels 1 and 1+ for factual questions and through level 3 for inferential questions. True-False-Not addressed questions discriminated most consistently between levels 2 and 2+. Cloze passages were powerful discriminators across all levels from 0+ to 3. The relative placement of the curves on the Y axis indicated that the easiest tasks were sign identification and information identification followed by multiple choice, True-False-Not Addressed, and Cloze.
An Evaluation Scale for Integrative Writing Tasks on a Performance Test of English for Academic Purposes

The Ontario Test of English as a second Language (OTESL) is a post-admissions test being developed for diagnosis and placement use in Ontario universities. Its primary purpose is to provide information on the examinee's ability to use English to perform typical academic tasks. The writing tasks on the test require the examinee to integrate information from the readings and lectures that precede the writing component. In devising the scoring system for these writing tasks we needed to develop a scale that told us more than just that the paper was well or poorly organized, though that, too, was important. In particular, we needed a scale that told us how well the examinee had succeeded in integrating the information, an ability we regard as essential for academic writing. In addition, we wanted to evaluate how appropriate the writing was for an academic audience and how well the examinee understood and carried out the rhetorical purpose of the task. As part of the test development project we compared the information provided by a variety of scoring schemes for our prompts. In this paper we report the results of that study and discuss the procedure we ultimately developed. In brief:

1) The information provided by analytic scoring schemes (Jacobs, et al., 1981; Brown & Bailey, 1984), while providing useful data about the general level of second language proficiency, did not tell much about the higher-level skills in which we were interested.

2) Primary trait schemes (Mullis, 1974) gave some information about the examinee's handling of the rhetorical situation, but it was too general and did not give much insight into the ability to integrate information.

3) A system developed by Faigley, et al., (1983) appears to offer an approach to measuring these higher-level skills. (An example is attached.) The difficulty with this scheme is that it is very task specific, making it difficult to generalize the results or to develop comparable tasks. It also requires very clearly defined writing prompts.

The writing scale being used in final trials of OTESL draws on all these systems. In this paper we report the results of using these different schemes with two writing tasks, allowing us to compare the different schemes within one task and the same scheme across different tasks.
Performance Profiles for Academic Writing

The British Council's English Language Testing Service uses an assessment scale to report overall performance on the test. It also uses more specific assessment scales to report performance on the two productive sections of the test, M2 (Writing) and M3 (Oral Interview). Each scale consists of nine bands with a 'performance description' of each band. The performance descriptions are intended to provide clear bases for raters to assign consistent ratings to language performance, written and oral.

This paper focuses on the written assessment scale for M2 (Writing). Since the assessment scale was written prior to the introduction of the test there have been a number of developments to improve the assessment of M2. In particular, the concept of profile reporting which is at the heart of the ELTS (that is, the view that a test candidate may not perform at the same level in, for example, listening as in writing, or in speaking as in reading) has been extended into the assessment of the academic writing section of the test. That is, it has been accepted that some writers may display, for example, excellent grammatical control but weak organizational skills. This view has been made explicit in the M2 Assessment Guide (developed by Liz Hamp-Lyons for the British Council/UCLES and in operation from March 1985). The assessment scale, however, is only now coming under review.

Clearly, if there is more than one criterion for the assessment of a piece of academic writing, each criterion needs to appear in the performance description. This is not a particularly difficult task. What is difficult, however, is to construct performance descriptions in such a way that they can handle not only unmarked profiles (that is, profiles which show the same level of performance on each criterion) but also marked profiles (that is, profiles which show distinct variations in performance on different criteria).

The paper will discuss these problems, expand on the rationale and evidence for the use of performance profiles for academic writing, and describe the development work which is being done on the M2 assessment scale.
Scalar Analysis of Composition Ratings

Use of independent performance ratings is common in the assessment of language proficiency in the productive skills of writing and speaking. Often the use of global inter-rater reliability estimates and holistic scales of measurement leaves unanswered questions about measurement accuracy and validity for individual examinees at specified points along the measurement scale.

The present study employs Rasch Model Microscale procedure in the analysis of 133 university-level ESL compositions of 75-minute duration. Through this procedure it is possible to examine rating accuracy at all points along each of the five different subscales. Use was made of a scale consisting of five, five-point rating subscales and two or three raters to rate compositions prepared over a uniform topic. The five subscales under investigation were content, organization, expression, structure and mechanics. Concurrent measures were available in the form of scores on a multiple-choice writing error detection test and actual placement levels in an instructional program. Factor analytic tests of data dimensionality are presented along with person and rating fit analyses to provide estimates of response validity that were not possible through traditional classical measurement analyses. Multiple regression analyses address the problem of relative weighting of the five different rating scales at various points along the proficiency continuum.

Recommendations are made regarding scalar refinement in the use of ratings with language production evaluation. Specific information is presented regarding the effectiveness of the five rating subscales and the five scale steps used on each subscale. Additional information on individual person fit is provided through carefully conducted interviews of misfitting examinees.
A pervasive question in the operational use and interpretation of the results of speaking proficiency interviews based on the "ILR" (Interagency Language Roundtable) proficiency level descriptions is the extent to which given examinees' performances are evaluated in a similar manner across the variety of government agencies and other institutions that make use of this testing procedure. This paper describes the procedures and major results of a direct experimental comparison of the speaking proficiency ratings assigned to the same group of examinees by testers in each of three government language training agencies (CIA, DLI, and FSI) for each of two languages (French and German). Based on a total of 61 examinees in French and 54 in German, who were interviewed sequentially over a two-day period by examiner teams from each of the three participating agencies, nonsignificant across-agency differences in proficiency level ratings on an overall basis were found for both combined and separate language groups. However, some agency-specific tendencies toward "generosity" or "severity" of ratings within certain portions of the overall scale were noted, as were occasional widely discrepant ratings of given examinees across the three agencies. In addition to the major statistical results, other related data are reported and discussed, including intercorrelations of the recently revised/defined "comprehension," "discourse," "structure," "lexicalization," and "fluency" factors, as evaluated by the FSI testers. The paper concludes with a summary of questionnaire-based information provided by testers and examinees concerning the perceived adequacy/fairness of the elicitation procedures used during the interviews, the extent to which affective aspects were properly handled, and so forth. Recommendations are made for follow-up studies including more detailed linguistic and elicitation procedure analysis of the over 300 tape recorded interviews obtained in the course of the project.
The Public Service Commission of the Government of Canada has adapted the ILR Interview for use in testing the second language proficiency (in English and French) of Canadian employees. The paper will describe how the ILR rating criteria, the rating scale, content and elicitation techniques were changed to adapt the test to the needs of the Canadian Government. This will include a discussion of the problems posed by administering the test to a large test population (20,000 tests annually) which is spread out over a wide geographic area and which includes trades, clerical, scientific, administrative, and executive personnel. The paper will also look at the selection and training process for test administrators, as well as at monitoring and on-going training activities. In making the presentation, overhead projection slides will be used to highlight the main points and to provide a framework for the talk.
The Graphic Representation of Language Competence: Mapping EFL Proficiency Using Multidimensional Scaling Techniques

Two important questions currently extant in the field of language ability testing are: (1) whether language proficiency is underpinned by a general unitary factor or by multiple factors and (2) whether the assumption of unidimensionality which underlies latent trait models has been seriously violated in a specific case. A further question which derives naturally from these two is what impact such a violation would have on a given test which has been shown to assess a multidimensional trait but whose items have been analyzed according to models which assume unidimensionality.

Heretofore, factor analysis (FA), both classical and, recently, confirmatory (cf. Vollmer, 1985), has been the statistical technique used to analyze the dimensionality of language tests. Although FA is widely used for this purpose, it has been open to criticism, particularly in its lack of parsimony in its explication of underlying factors. In part because of the drawbacks of FA, and in part in an effort to graphically represent the relationship among the test items and thus the competence of the learners, in this study another statistical technique was used, specifically, multidimensional scaling (MDS).

This technique, which is in some ways parallel to factor analysis, does not require the investigator to assume that the factors are related in a linear fashion, as FA does. In addition, MDS has the potential to be more parsimonious than FA. Also, since it is basically a spatial technique, it can provide us with a graphic representation of the language competence of the learners at a given time.

Although it has been used in a wide variety of ways in other fields, the study which will be described in this paper represents the first application of this technique in the language testing field. The data used in the analysis, which includes the responses of approximately 2000 basic-level EFL adult learners to criterion-referenced tests, were processed as follows:

1) Rasch (BICAL) analyses of the items were produced;
2) inter-item product moment correlation matrices were constructed;
3) the matrices were analyzed using the KYST (MDS) program;
4) the outputs of the KYST and Rasch analyses were assessed separately and then were juxtaposed.

Conclusions were then drawn concerning the dimensionality of the test items, as well as the reliability and validity of MDS as an alternative to FA.
Comparison of Two Forms of Tutor Assessment

As a means of establishing the predictive validity of ELTS, a number of procedures have been developed to monitor students' progress after completing the recommended language training and starting their courses of academic study. The first procedure involves tutor assessment at the end of the recommended language training. A special form has been devised for this assessment: Test Validation Form 1. A second procedure involves the use of Test Validation Form 2, a form completed by subject tutors in academic institutions.

These forms were devised to facilitate the encoding of the assessment on a computer program, which in turn will allow easy access to and processing of data accumulated over a period of time. The data are constantly being added to, with the result that the statistical base for any analysis is steadily growing larger.

Here I am going to look at Test Validation Form 1 only and describe a small-scale investigation. The form requires the tutors in language schools to assess a student's standard of English in two ways. The first is according to four categories of general assessment (i.e., fluent, good, fair, and weak) and the second is according to ELTS Band Descriptors, 0 through 9. Assessments are made for each of the four skills (listening, speaking, reading and writing) separately. This double assessment was originally devised as a check on the consistency of tutor assessments. Now that a considerable number of Test Validation Form 1 have been returned and fed into the computer, it is possible to compare the assessments according to the general 'lay' categories with those made by the same tutors according to ELTS bands.
An Experiment with an On-Line Narrative Discourse Test

Recent trends in foreign language proficiency testing have turned to the oral interview as the preferred method of evaluating the speaking skills of foreign language learners. The time and personnel costs of the oral interview, however, place severe constraints on its applicability in large-scale foreign language programs. An alternative to the face-to-face DLI type interview is the on-line narrative discourse test which can be administered en masse to foreign language learners at any level in a language laboratory equipped with video facilities.

The present study outlines the construction of a video taped on-line discourse test made to evaluate the speaking skills of O+ level Japanese EFL learners. Methods of establishing inter-rater reliability and scaling procedures will be discussed in the presentation. Concurrent validity with a more time-consuming structural interview method will be discussed along with data derived from factor analyses of eleven subjective and objective subscores used in the scoring of each learner's narrative. Methods used in the derivation of measures incorporating both aspects of fluency and accuracy in speaking skill will be presented in the discussion of the factor structure of the narrative discourse. Component scores used in the analysis include holistic ratings of pronunciation, fluency, accuracy and a total score. The objective measures include total spoken t-units, error-free t-units, and the mean number of words in error-free t-units.
Ten Parameters for Measuring Pronunciation

Current measurement scales of language proficiency treat pronunciation in a subjective manner. Rather than referring to specific features such as the degree of insufficient voicing, they refer to general impressions, such as comprehensibility. This contributes to the wide and often conflicting range of evaluations which has caused so much confusion regarding the proficiency of foreign teaching assistants.

State of the art computers designed for speech processing provide reliable and efficient means for measuring the specific features which contribute to comprehensibility. These features are voicing, aspiration, turbulence, frequency, missing sounds, extra sounds, duration, stress, transitions, and volume. Measurement of these features is essential for standardizing characterizations of the pronunciation of non-native speakers of American English.

This paper presents results of research on establishing measurement scales of the pronunciation of 50 international graduate students from a variety of third world nations. They were all participants in an experimental course, Pronunciation Lab.

Equipment utilized in this study includes a Kay Elemetrics Speech Spectrographic Display (SSD 8800), a Kay Elemetrics Visi-Pitch (SE 6095), an Apple IIe PC, a VCR, tape deck, and oscilloscope.

Other aspects of the research project include computer assisted instruction in pronunciation and the characterization of foreign accent.
Innovation Through Computer-Based Language Testing (CBELT)

It has been occasionally claimed that the use of computers in language testing will not increase the validity of language tests, and that while the computer may help us to do more efficiently what we can already do without it, it will not lead to innovation in the type of tests that can be developed. Indeed, it has been argued that the use of the computer to deliver language tests will have a negative, restraining effect on development in so far as it will lead to a continued reliance on closed choice item types (i.e., multiple choice and cloze tests.) Certainly recent attention in language testing circles has focussed on the development of statistical treatments of test data and of adaptive test delivery systems, rather than on the nature of the test items and the test content.

The aim of the research reported on in this paper was to challenge the above claims and to explore the possibility for innovation in test item types that might exist when language tests are delivered by computer. The paper will outline the main findings of the exploratory research, which appeared to suggest that previous pronouncements were unduly negative. The research has identified promising areas for focus and development in CBELT, which might lead to improvements in test types and possibly to increases in test validity.

**Error analysis — Burgess**

**Branching for presentation of items on basis of error pattern.**

**Input / Stimulus**

**Response types**
Some Problems Concerning the Interpretation of Passage Correction Tests

Passage Correction (PC) tests, which require individuals to identify and correct errors that have been inserted in a prose passage, have been shown to correlate significantly with other EFL tests (e.g., Davies 1975G, Arthur 1980). These and other studies suggest that PCs may be valuable for research in second language acquisition, and testers at several universities have adopted, or have considered adopting, PCs for their programs. However, the characteristics of such tests are much less understood than those of cloze tests and other measures. The aim of this paper is to examine some PCs that have some of the statistical characteristics of reliable and valid language tests but that call for close scrutiny in three areas:

1) The use of native-speaker performance as a baseline. While using native-speaker performance as a baseline is often a justifiable procedure, there are problems in doing so on some PCs. An error that has been inserted may sometimes be less frequently detected (or corrected) than some parts of the passage as they were originally written.

2) Individual variation among EFL students on the PC. While both holistic essay evaluations and PCs often correlate significantly with other EFL measures there are cases where individuals perform much better or worse on PCs than one might predict on the basis of holistic essay evaluation.

3) Relationships among individual PC items. Despite the high measures of internal consistency that some PCs show, there are common instances of where two items involving the same type of error do not correlate at all. Nevertheless, multiple regression analyses show that other items collectively can usually predict performance on PC items.

Testing strategic competence - metalinguistic awareness

Questions
1) Procedure / criteria for selecting items to be analyzed.
Identification of target items.

2)
Utilizing Rasch Analysis to Detect Cheating on Language Exams

The purpose of this presentation is to demonstrate the feasibility of utilizing computerized Rasch analysis to detect cheating on objective language exams such as the TOEFL.

The Rasch on-parameter logistic model (Rasch 1980) has been successfully utilized for such varied purposes as precise test-item calibration (Englehard 1980), learner-specific item analysis (Rasch 1966), item banking (Rubin and Mott 1983), test linking, as well as computerized testing (Henning 1984), and identification of item bias (Larson and Madsen 1985).

Computerized Rasch applications in testing are timely given concerns that have been expressed both overseas (Roizen 1982) and in the United States (Cziklo and Lin 1984). Its pervasiveness is illustrated by various reports (Singhal and Johnson 1983, Nuss 1984), and its causes by a variety of studies (Houston 1983, Antion and Michael 1983, Guttmann 1984; Forsyth and others 1985).

The investigation reported in this paper utilizes Wright and Linacre's Microscale analysis developed for the IBM pc. A pre-study of contrived cheating data from 250 actual ESL language exams demonstrates the robustness of the program in identifying cheaters as "outliers" on the Student Outfit plot on grammar, listening, and reading subtests. (Students range in ability from beginning to intermediate level.) In the principal study, three groups of 30 TOEFL scores are Rasch analyzed: each group containing one apprehended cheater. ETS arranged to conceal the identity of each of the three cheaters until after the study was completed. The procedure followed in the principal investigation concentrates on identifying those paid cheaters who have arbitrarily suppressed their scores in order to provide their "client" with a pass that would not be noticeably high.
Roundtable 4: Composition of various language testing programs